Candidate Identification Risk Certification Report

Yannelly Mercado
3/10/23

Executive Summary

-The risk of patient identification was evaluated for the candidate report “survey_data.xls”

- I applied different methods in order to calculate the risk identification. These methods included probabilistic algorithms and the review of risk under different scenarios.

- Recommendations for reducing the identification risk can be found on appendix 1

- A list of different problems relating to the data can be found in appendix 2

Data Description

- Customer Survey Data
  - 4,329 Customers
  - Average age: 50
  - Average household income: 58,000
  - Average employment length: 12
  - Average debt to income ratio: 9.9

- Timeframe: 7/11/2021 - 5/9/2022

- Variables(columns): Gender, Age, Education Years, Household Income, Debt To Income Ratio, Employment Length.

Data Description

- Each customer age ranges from 18 to 79 years
- Each customer has education years ranging from 6 to 23 years
- Each customer has a household income ranging from 9 to 1000 dollars
- Each customer has a debt to income ratio between 0 to 43
- Each customer has an employment length ranging between 0 and 52 years
- The data was divided into subgroups of 2, 6, and 15. With 15 being the last grouping.

Customer Risk Identification Methodology

- Step 1: Quality control: The data was thoroughly analyzed and examined for outliers, extreme low/high values, and null values. Please refer to appendix 2 for issues related to the data.
- Step 2: Quasi-identifiers (customer indirect variables): Gender, Age, Education Years, Household Income, Debt To Income Ratio, Employment Length. I have not considered customer ID in this report since this is a variable more suitable for re-identification.
- Step 3: Equivalence classes: these were created from the data by grouping together a certain amount of records with the same combination of values and indirect variables.

Customer Risk Identification Methodology

Step 4: After many iterations, and considering the sensitivity of the data, we identified that the smallest equivalence class has a size of at least 2 customers. This means that the threshold risk is about 0.5(50%). For more information about risk thresholds relating to banks, please refer to Bank for International Settlements.

Customer Risk Identification Methodology

- Step 5: Four different attack scenarios were considered:
  - Scenario 1: An intruder attempts to purposely identify customer data
  - Scenario 2: An intruder inadvertently identifies customer data
  - Scenario 3: Data is exposed by data breach at recipient's site
  - Scenario 4: An enemy launches an attack to demonstrate data vulnerability. The goal is to show the data can be re-identified

Scenario 1: Deliberate data attack

- I assumed that someone in the recipient's site tried to identify the data, on purpose. The purpose is unknown, but can include monetary reasons, or snooping.
- Calculation process:
  - Probability:
    - 300 employees with access to the data
    - 60 employees go rogue
- Probability of attempt is 60/300 = 0.2 50%
- This probability was computed using the data from the dataset

Scenario 2: Inadvertent Data Attack

- I assumed that someone at the recipient's site recognized the customer in the dataset. For example, a loan officer working with the dataset and recognizes his/her family member/neighbor.
- Calculation process:
  - Probability:
    - 27.70 % bank approval rate
    - 150 average number of friends people have
- Probability of attempt is 1-(1-0.02770) ^150 99%
- This probability was computed using the data from the dataset
- Resource used to obtain the bank approval rate is Fundera.com

Scenario 3: Data Breach

- I assumed that the data recipient lost the dataset. For example, the recipient's credentials were stolen.
- Calculation process:
  - Probability:
    - 19% Average percent of people who get their credentials stolen.
- Probability of breach is 19%
- This probability was computed using the data from the dataset
- Resource used to obtain the percentage of credentials is IBM's data breach report

Scenario 4: Demonstration Attack

- I assumed that the data was shared publicly (a person has enough private information to launch an attack on the data and attempts to re-identify the customers).
- If the person is known, it can be identified through scenarios 1-3.
- In this scenario, the person is unknown. As we can’t identify all of the people who will be using the dataset.
- Probability of risk: 100%
- This probability was computed using the data from the dataset

Results - Diagnostics

- I assumed that the data recipient lost the dataset. For example, the recipient's credentials were stolen.
- Calculation process:
  - Probability:
    - 19% Average percent of people who get their credentials stolen.
- Probability of breach is 19%
- This probability was computed using the data from the dataset
- Resource used to obtain the percentage of credentials is IBM's data breach report

Results - conclusion

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Max risk	50%	99%	19%	100%
Median risk	33%			
Assessment	Unacceptable	Unacceptable	Acceptable	Unacceptable

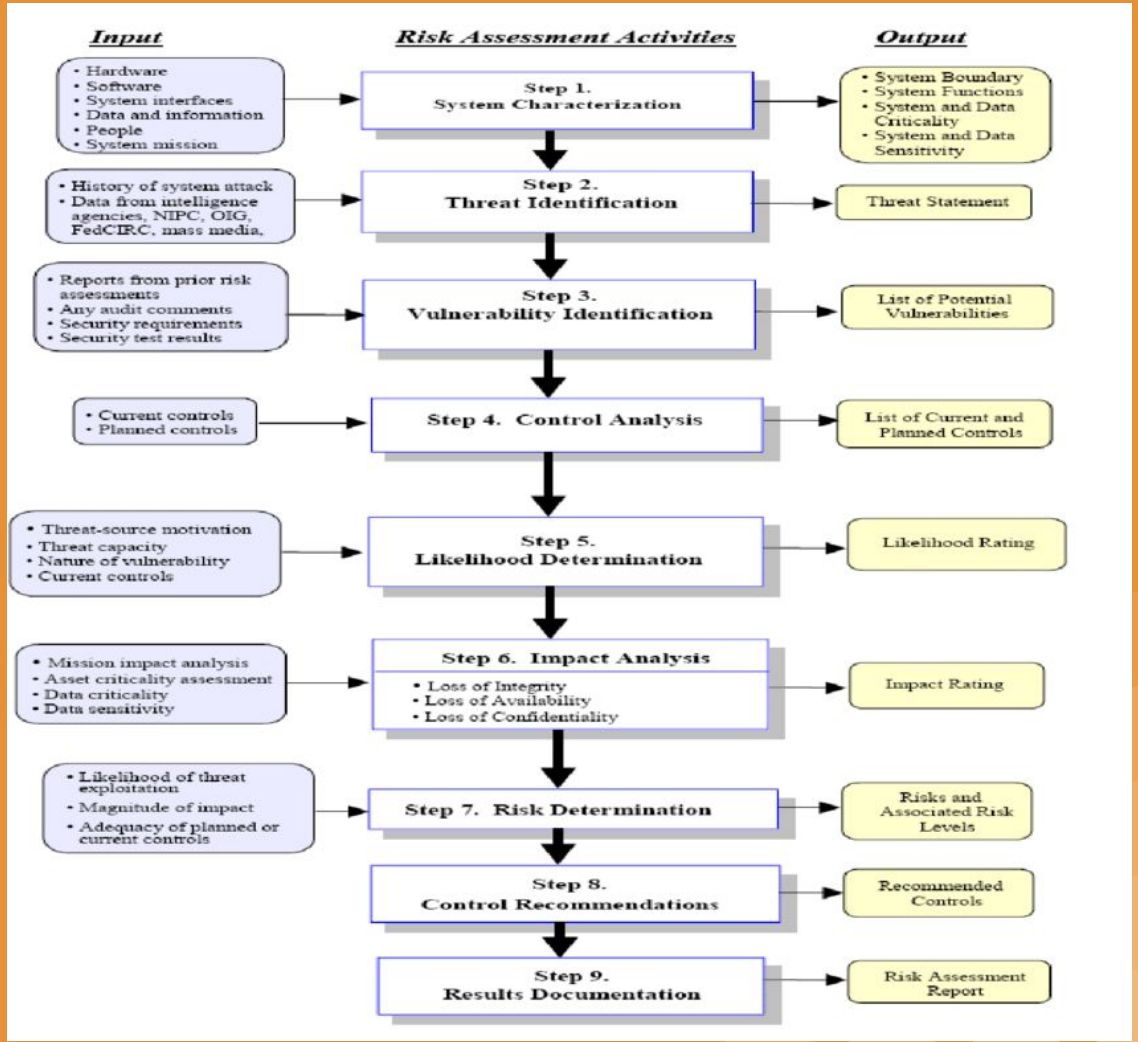
Appendix 1: Recommendations

- To lower the risk of re-identification for the dataset:
  - Mask the data
  - Identify outliers and decide how to move on
  - Group education years
  - Group ages
  - Group household income
  - Group in ranges of 6 or 15

Appendix 2: Quality Control

- Outliers:
  - One individual with a 1073 household income (out of range compared to others. Household income was an average of 54)
  - 3 individuals with a debt to income ratio greater than 40 and 1 with less than 0.1 (average ratio was 10, max was 43)
  - 5 individuals with employment length above 45 (average length was 10 and max value was 52)

Risk Assessment Methodology Recommended



Reference: Towards a Risk Assessment and Evaluation Model for Economic Intelligent Systems

References

Basel Committee on Banking Supervision Consultative Document Guidelines. https://www.bis.org/bcbs/publ/d398.pdf

"Cost of a Data Breach 2022." *IBM*, https://www.ibm.com/reports/data-breach?utm_content=SRCWW&p1=Search&p4=43700072379268724&p5=p&gclid=CjwKCAiAjPyfBhBMEiwAB2CCIsX8n31iNCT5XQbRg6IFte3ohXSURwF86uA5mFSlkIGRweg5WIisrRoCdSYQAvD_BwE&gclsrc=aw.ds

Shepherd, Maddie. "Small Business Lending Statistics and Trends." *Fundera Ledger*, Fundera, 23 Jan. 2023, https://www.fundera.com/resources/small-business-lending-statistics#:~:text=Alternative%20lenders%20have%20a%2056.8,have%20a%2027.7%25%20approval%20rate

Towards a Risk Assessment and Evaluation Model for Economic Intelligent Systems https://www.researchgate.net/publication/341051987_TOWARDS_A_RISK_ASSESSMENT_AND_EVALUATION_MODEL_FOR_ECONOMIC_INTELLIGENT_SYSTEMS